DOCKET NO. SAMS01-00149

PATENT

A NEW MECHANISM FOR CELL ROUTING IN A MULTI-STAGE FABRIC WITH INPUT QUEUING

Inventor(s)

Sreedharan P. Sreejith
205 Misty Glen Lane
Murphy
Collin County
Texas 75094
Citizen of India

Praseeth K. Sreedharan
136 Prairie View
Murphy
Collin County
Texas 75094
Citizen of India

Assignee:

SAMSUNG ELECTRONICS CO., LTD.
416, Maetan-dong, Paldal-gu
Suwon-city, Kyungki-do
Republic of Korea

CERTIFICATE OF EXPRESS MAIL

I hereby certify that this correspondence, including the attachments listed, is being deposited in an envelope addressed to the Assistant Commissioner of Patents, Washington, DC 20231 as "Express Mail, Post Office to Addressee" on the date indicated below.

Kathy Longenecker
Printed Name of Person Mailing

Kathy Longenecker
Signature of Person Mailing

ET 649646900 US
Express Mail Label No.

8/20/01
Date

William A. Munck
John T. Mockler
NOVAKOV, DAVIS & MUNCK, P.C.
P.O. Drawer 800889
Dallas, Texas 75380
(214) 922-9221

A NEW MECHANISM FOR CELL ROUTING IN A MULTI-STAGE FABRIC WITH INPUT QUEUING

TECHNICAL FIELD OF THE INVENTION

The present invention is directed, in general, to routing traffic through a multi-stage switch mesh or a switch fabric having multiple paths between each pair of input and output ports and, more specifically, to selecting cell paths within such a multi-stage switch mesh or multi-path switch fabric employing input queue to minimize head of line blocking and improve throughput for large numbers of independent traffic flows.

BACKGROUND OF THE INVENTION

A switch (or "switch fabric") routes data traffic from one of N input ports to one of N output ports. A multi-stage switch mesh includes a plurality of switches inter-connects via inputs and outputs to provide a non-blocking architecture effectively acting like a large switch fabric with a large number of input and output ports.

In connection-oriented (as opposed to packet-oriented) technologies such as asynchronous transfer mode (ATM), a cell path is computed from an input port through the switch or switch mesh to an output port during the connection setup time and remains the same throughout the lifetime of the connection (e.g., until a requested data transfer is complete). The path is uniquely identified by a label and each cell presented at the input port of the first switch within the path has this routing label attached, together with an assigned or requested priority. The switch fabric or switch mesh fabric places queues received cells in the input queues of the corresponding priority for the input port identified in the routing label, where each input port has associated therewith more than one input queue each having a different priority. The cell scheduler within the switch fabric identifies the output port from the routing label and transfers the cell based on the associated priority.

The input queues of a switch fabric are typically of a fixed size, with one input queue for each possible priority associated with each input port. Each port and each input queue is normally independent, with no sharing of resources. If for some reason an input queue becomes full and cannot receive any more cells, the upstream traffic source is informed utilizing a flow control mechanism such as "back-pressure," which effectively reduces the congestion in the forward direction by blocking the traffic at the source itself.

Input queues become congested when cells cannot be placed to the output port because higher or equal priority queues from the same input port or a different input port are scheduled for transfer to the desired output port before the subject cell. Cell departure must therefore be delayed or rescheduled for a later time.

One effect of cell congestion is "head of line" blocking, where cells queued at an input port cannot be serviced even if the corresponding output port is free because another cell ahead of the blocked cell(s) is still waiting for resources from a different output port. Congestion thus spreads very quickly within the system and reduces overall switching throughput.

In a multi-stage switch mesh, or a switch fabric including multiple paths between each input and output port, the overall performance depends heavily on the traffic path of various cell streams through the switch or switch mesh. There is, therefore, a need in the art for a method of selecting cell flows with switch fabric or a switch mesh to minimize head of line blocking and improve overall performance.

SUMMARY OF THE INVENTION

To address the above-discussed deficiencies of the prior art, it is a primary object of the present invention to provide, for use in selecting a path for traffic flow within a multi-stage switch mesh, a head of line (HOL) blocking count value, which is directly proportional to the committed traffic load through an output port and is computed for a path by adding the values associated with all output ports within the path. In selecting a route, all paths from the source to the destination are identified and sorted by head of line blocking count value. Rather than selecting a path based on traffic load, the path having the lowest head of line blocking count value and sufficient capacity for the requested traffic is selected, with selection between paths having equal head of line blocking count values being made based on traffic load.

The foregoing has outlined rather broadly the features and technical advantages of the present invention so that those skilled in the art may better understand the detailed description of the invention that follows. Additional features and advantages of the invention will be described hereinafter that form the subject of the claims of the invention. Those skilled in the art will appreciate that they may readily use the conception and the specific embodiment disclosed as a basis for modifying or designing other structures for carrying out the same purposes of the present invention. Those skilled in the art will also realize that such equivalent constructions do not depart from the spirit and scope of the invention in its broadest form.

Before undertaking the DETAILED DESCRIPTION OF THE INVENTION below, it may be advantageous to set forth definitions of certain words or phrases used throughout this patent document: the terms "include" and "comprise," as well as derivatives thereof, mean inclusion without limitation; the term "or" is inclusive, meaning and/or; the phrases "associated with" and "associated therewith," as well as derivatives thereof, may mean to include, be included within, interconnect with, contain, be contained within, connect to or with, couple to or with, be communicable with, cooperate with, interleave, juxtapose, be proximate to, be bound to or with, have, have a property of, or the like; and the term "controller" means any device, system or part thereof that controls at least one operation, whether such a device is implemented in hardware, firmware, software or some combination of at least two of the same. It should be noted that the functionality associated with any particular controller may be centralized or distributed, whether locally or remotely. Definitions for certain words and phrases are provided throughout this patent document, and those of ordinary skill in the art will understand that such definitions apply in many, if not most, instances to prior as well as future uses of such defined words and phrases.

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the present invention, and the advantages thereof, reference is now made to the following descriptions taken in conjunction with the accompanying drawings, wherein like numbers designate like objects, and in which:

FIGURES 1A and 1B depict a switch and a switch mesh, respectively, in which cell flows are selected to minimize head of line blocking according to one embodiment of the present invention; and

FIGURES 2A and 2B are high level flowcharts for a process of selecting communications paths for requested cell traffic through a multi-stage switch mesh according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

FIGURES 1A-1B and 2A-2B, discussed below, and the various embodiment used to describe the principles of the present invention in this patent document are by way of illustration only and should not be construed in any way to limit the scope of the invention. Those skilled in the art will understand that the principles of the present invention may be implemented in any suitably arranged device.

FIGURES 1A and 1B depict a switch and a switch mesh, respectively, in which cell flows are selected to minimize head of line blocking according to one embodiment of the present invention. FIGURE 1A depicts a switch fabric 100 including a plurality of input ports 101a-101n (where "n" represents any positive, nonzero integer) and a plurality of output ports 102a-102n. Each input ports 101a-101n has associated therewith one of a number of sets of input queues 103a-103n, where each set of input queues 103a-103n includes a queue for each possible priority. Each output port 102a-102n has associated therewith a scheduler 104a-104n for scheduling cell transfer from a queue within one of the sets of input queues 103a-103n to the respective output port 104a-104n. Switch fabric 100 may include redundant communications paths (i.e., multiple parallel paths) between each input port 101a-101n and/or the associated input queues 103a-103n to each of the output ports 102a-102n.

FIGURE 1B depicts a switch mesh 105 including an array of interconnected switches 100aa-100am, 100ba-100bm, and 100la-100lm (where "l" and "m" are any positive, nonzero integers, not necessarily equal, although "m" preferably equals "n"), each switch having the structure of switch 100 depicted in FIGURE 1A. Switch mesh 105 is organized in a plurality of stages 100a, 100b and 100l, preferably three. The input ports 106 of switch mesh 105 are the input ports of the switches 100aa-100am within the first stage 100a, while the output ports 107 of switch mesh 105 are the output ports of the switches 100la-100lm within the last stage 100l. As illustrated, the remaining input and output ports of switches 100aa-100am, 100ba-100bm, and 100la-100lm are interconnected in accordance with the known art to allow switch mesh 105 to operate as a large or extended switch fabric.

FIGURES 1A and 1B and the above description do not depict or describe all details of the complete construction and operation of switch 100 and switch mesh 105. However, those skilled in the art will understand that the present invention may be practiced with conventional switches or switch meshes, and only so much of the construction and operation of such conventional switches or switch meshes which is unique to the present invention or necessary for an understanding of the present is depicted in the figures and described herein.

Input queues 103a-103n and/or schedulers 104a-104n within switch 100, and within each switch 100aa-100lm within switch mesh 105, are coupled to a cell routing server 108, a centralized controller for determining cell routing through the switch 100 or switch mesh 105 and selecting the cell path from the input port receiving the subject cell to the target output port.

The traditional approach for selecting cell paths through a switch or switch mesh is based on selection of a "least loaded path" between the input ports and output ports. Although this approach provides sufficient throughput under normal circumstances, switch fabrics or meshes which employ input queuing and flow control based on back pressure can experience congestion as described above, with throughput worsening as the number of independent traffic flows through the switch fabric or mesh increases.

Selection of paths according to the present invention follows a two-dimensional approach where both head of line blocking effects and traffic load are considered for path calculation. Path selection by cell routing server 108 in the present invention relies on two assumptions:

1. As long as the traffic load in within the physical capacity of the switch fabric or switch mesh, traffic load does not affect performance; and

2. Interference of traffic streams from other input ports (i.e., head of line blocking) is the principal factor affecting overall switch performance.

Cell routing server 108 is linked to connections table or database 109 containing an identification of all currently active path connections from input port(s) to output port(s) through switch 100 or switch mesh 105, including bandwidth statistics. For each output port 102a-102n or 107 of switch 100 or switch mesh 105, cell routing controller 108 maintains a used capacity count 110, the cell stream data transfer rate expressed in cells/second, and head of line blocking counts 111 for each priority in that port.

For each output port 102a-102n or 107, one input port is designated as the "desired" input port, with traffic from that input port 101a-101n or 106 having a head of line blocking count value of zero (i.e., no effect). All remaining input ports are considered "undesired" input ports for the subject output port. The head of line (HOL) blocking count value for a traffic flow is directly proportional to the committed traffic load (e.g., equal to the traffic load in cells/second).

When a connection to an output port from an undesired input port is added, the head of line count for the new traffic flow is added to the current sum of head of line counts for the corresponding priority at the target output port. Thus, connections table 109 will contain a plurality of head of line blocking count values for each output port 102a-102n or 107, one for each possible priority. The head of line blocking count value for a given priority is the sum of all head of line blocking counts for the existing connections to the subject output port with higher or equal priorities.

Thus, if a given output port 102n is, at the moment of interest, the subject of connection paths A, B and C having committed remaining traffic loads of C_A, C_B and C_C

cells/second, respectively, where each of connection paths A, B and C have a priority higher than or equal to a particular priority X, the head of line blocking count value for that particular priority X at the output port 102n is the sum of the head of line blocking count values of connection paths A, B and C, which equals to

$$HOL_{total} = HOL_A + HOL_B + HOL_C$$

Where:

$$HOL_A = K * C_A / C_L \text{ and}$$

$$HOL_B = K * C_B / C_L \text{ and}$$

$$HOL_C = K * C_C / C_L$$

Again, the term C_L is the total capacity (bandwidth) of the outgoing port. The term K is a variable whose value is 0 for connections from the desired ports, and is a system wide constant greater than 0 for connections from undesired input ports. (Note: the value of K should be selected in such a way that when large numbers of connections are added to the port, the counters holding the HOL values do not overflow. However, giving very small values for K may cause less accuracy during integer divisions in some systems).

For a single switch having multiple paths between each pair of input and output ports, the head of line blocking count is determined individually for each path. For a multi-stage switch mesh, the head of line blocking count for any path between an input port and an output port is the sum of the head of line blocking counts for all output ports within that path. Thus, for example, the head of line blocking counts for various paths from input port IN_AAA to output port OUT_LMN within multi-stage switch mesh 105 in FIGURE 1B are the sums of the head of line blocking counts for output ports: OUT_AAA, OUT_BAN and OUT_LMN; OUT_AAB, OUT_BBN and OUT_LMN; or OUT_AAN, OUT_BMN and OUT_LMN. Additionally, the head of line blocking count for a path at a particular priority is the sum of the head of line blocking counts at the respective priority for all output ports within that path.

In selecting a path for a cell stream, the optimal route is identified by cell routing server 108 by identifying all routes from the source input port to the destination output port and sorting those routes in order of the associated head of line blocking counts. The path having the smallest head of line blocking count--and sufficient capacity for the requested cell traffic in addition to existing traffic flows within that path--is selected. If more that one path has the same head of line blocking count and sufficient capacity, the path with the least traffic load is selected.

FIGURES 2A and 2B are high level flowcharts for a process of selecting communications paths for requested cell traffic through a multi-stage switch mesh according to one embodiment of the present invention. FIGURE 2A illustrates a process of selecting a communications path. The process 200 begins with a request for routing of a new cell stream path from an input port to an output port (step 201). All paths between the source and destination are then identified and sorted by head of line (HOL) blocking count (step 202). From the sorted paths, the path or paths having the lowest head of line blocking count and sufficient capacity to handle the requested traffic load is identified (step 203). If only one path has the lowest head of line blocking count and sufficient capacity (step 204), the identified path is selected and the traffic routed along that path. If more than one path has the (same) lowest head of line blocking count and sufficient capacity for the requested traffic (step 204), the path is selected from among that group based on traffic loading of the various paths.

In either case, the head of line blocking count(s) for any path(s) affected by the newly added routing are updated (step 207) and the process becomes idle (step 208) until another request for routing traffic is received.

FIGURE 2B illustrates a process 209 of handling completion of a cell transfer (step 210). The head of line blocking count(s) for any path(s) affected by completion of the cell transfer are updated (step 211) and the process becomes idle (step 212) until cell transfer is completed.

The present invention provides improved overall throughput in multi-stage switch fabrics, or individual switches, which includes multiple paths between each input and output port and employs input queuing with flow control based on back-pressure. The present invention can also support priority queues and may therefore be easily adapted to support various service levels or categories.

It is important to note that while the present invention has been described in the context of a fully functional switch or multi-stage switch mesh, those skilled in the art will appreciate that the mechanism of the present invention is capable of being implemented and distributed in the form of a computer usable medium of instructions in a variety of forms, and that the present invention applies equally regardless of the particular type of signal bearing medium is used to carry out the distribution. Examples of suitable computer usable mediums include: nonvolatile, hard-coded or programmable type mediums such as read only memories (ROMs) or erasable, electrically programmable read only memories (EEPROMs), recordable type mediums such as floppy disks, hard disk drives, and read/write (R/W) compact disc read only memories (CD-ROMs) or digital versatile discs (DVDs), and transmission type mediums such as digital and analog communications links.

Although the present invention has been described in detail, those skilled in the art will understand that various changes, substitutions, and alterations herein may be made without departing from the spirit and scope of the invention it its broadest form.